

# (12) UK Patent Application (19) GB (11) 2 366 633 (13) A

(43) Date of A Publication 13.03.2002

(21) Application No 0021881.8

(22) Date of Filing 06.09.2000

(71) Applicant(s)

Argo Interactive Group Plc  
(Incorporated in the United Kingdom)  
Oak House, Shackleford Road, ELSTEAD, Surrey,  
GUB 6LB, United Kingdom

(72) Inventor(s)

Roger Ian Spooner

(74) Agent and/or Address for Service

D Young & Co  
21 New Fetter Lane, LONDON, EC4A 1DA,  
United Kingdom

(51) INT CL<sup>7</sup>

G06F 17/21 // G06F 17/30, H04L 29/06

(52) UK CL (Edition T )

G4A AUXX

(56) Documents Cited

EP 0994426 A2 EP 0969389 A2

Wireless-adaption of WWW content over CDMA.Ham  
et al.Mobile Multimedia Communications '99 p.  
368-372

(58) Field of Search

UK CL (Edition S ) G4A AUDB AUXX

INT CL<sup>7</sup> G06F 17/21 17/30, H04L 29/06

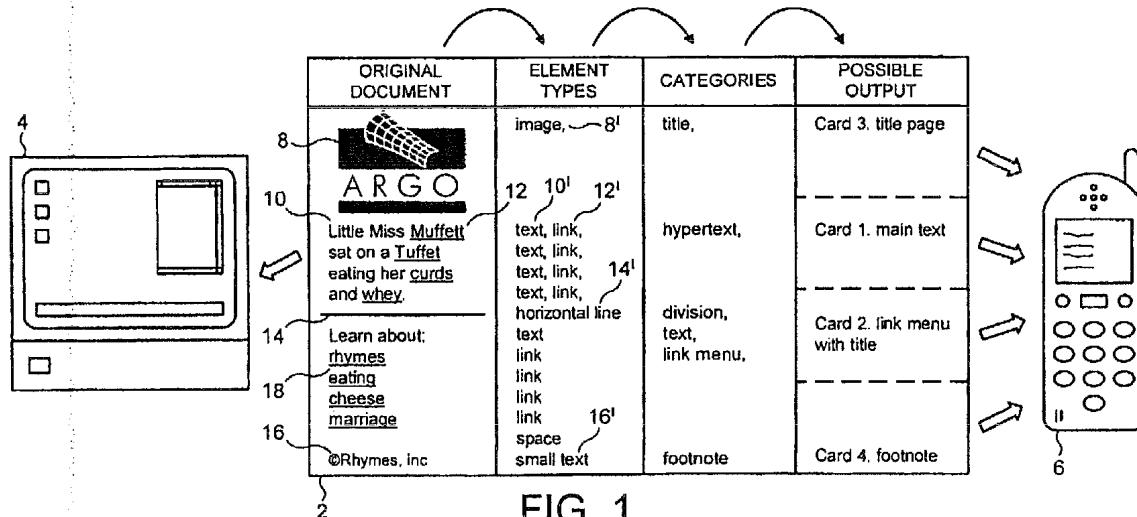
Online: EPODOC, WPI, PAJ, IEL

(54) Abstract Title

Analysing hypertext documents

(57) Hypertext documents are parsed so as to identify regions of text within the document. Each region of text is formed from a plurality of document elements which are found whilst parsing the document. These element types are categorised in regions and these categorised regions are identified dependant on a confidence measure associated with the regions.

One embodiment of the invention uses heuristic pattern matching on the categorised element types to identify patterns indicative of the different regions within the document. The original document may then be divided into separate documents based on the identified document portions. This method can be used for HTML, WML, CHTML or PDF documents.



GB 2 366 633 A

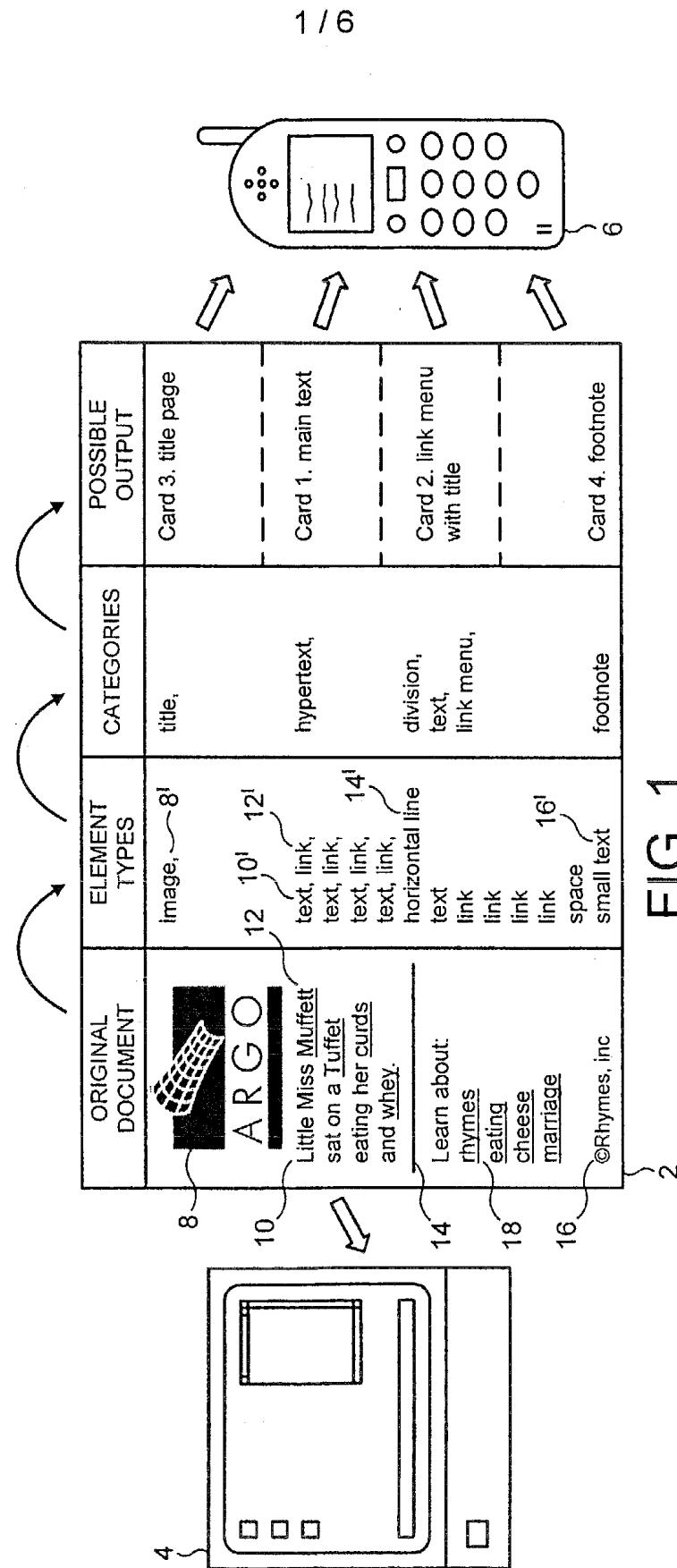


FIG. 1

05 + 3 01

2 / 6

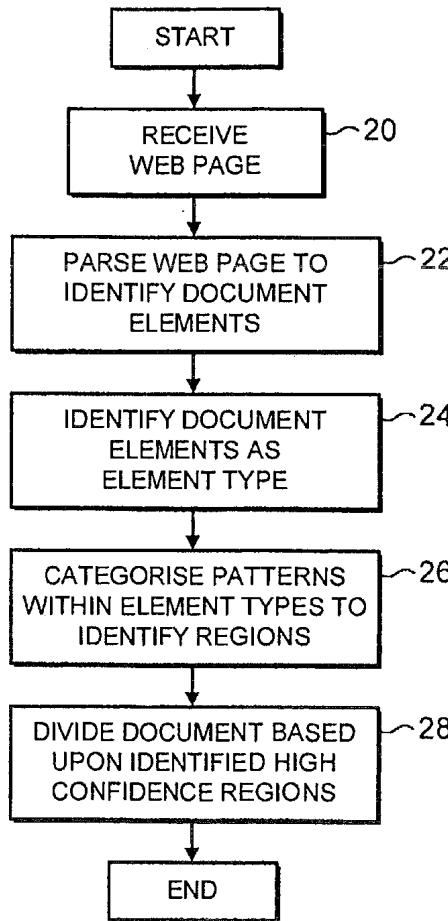


FIG. 2

	<u>ELEMENT TYPE(S)</u>	<u>ELEMENT(S) POSITION</u>	<u>REGION</u>
A.	MIXED TEXT AND LINKS	ANYWHERE	HYPertext BLOCK
B.	NEIGHBOURING LINKS	ANYWHERE	LINK MENU
C.	TEXT	IMMEDIATELY ABOVE A LINK MENU	LINK MENU TITLE
D.	IMAGE	CLOSE TO TOP	TITLE
E.	HORIZONTAL LINE	SPACED FROM ANY IMAGES OR GRAPHICS	DIVISION
F.	SMALL TEXT	CLOSE TO BOTTOM OR TITLE	FOOTNOTE
G.	IMAGE	SURROUNDED BY HORIZONTAL LINES	IMAGE

FIG. 3A

05 + 3 01

3 / 6

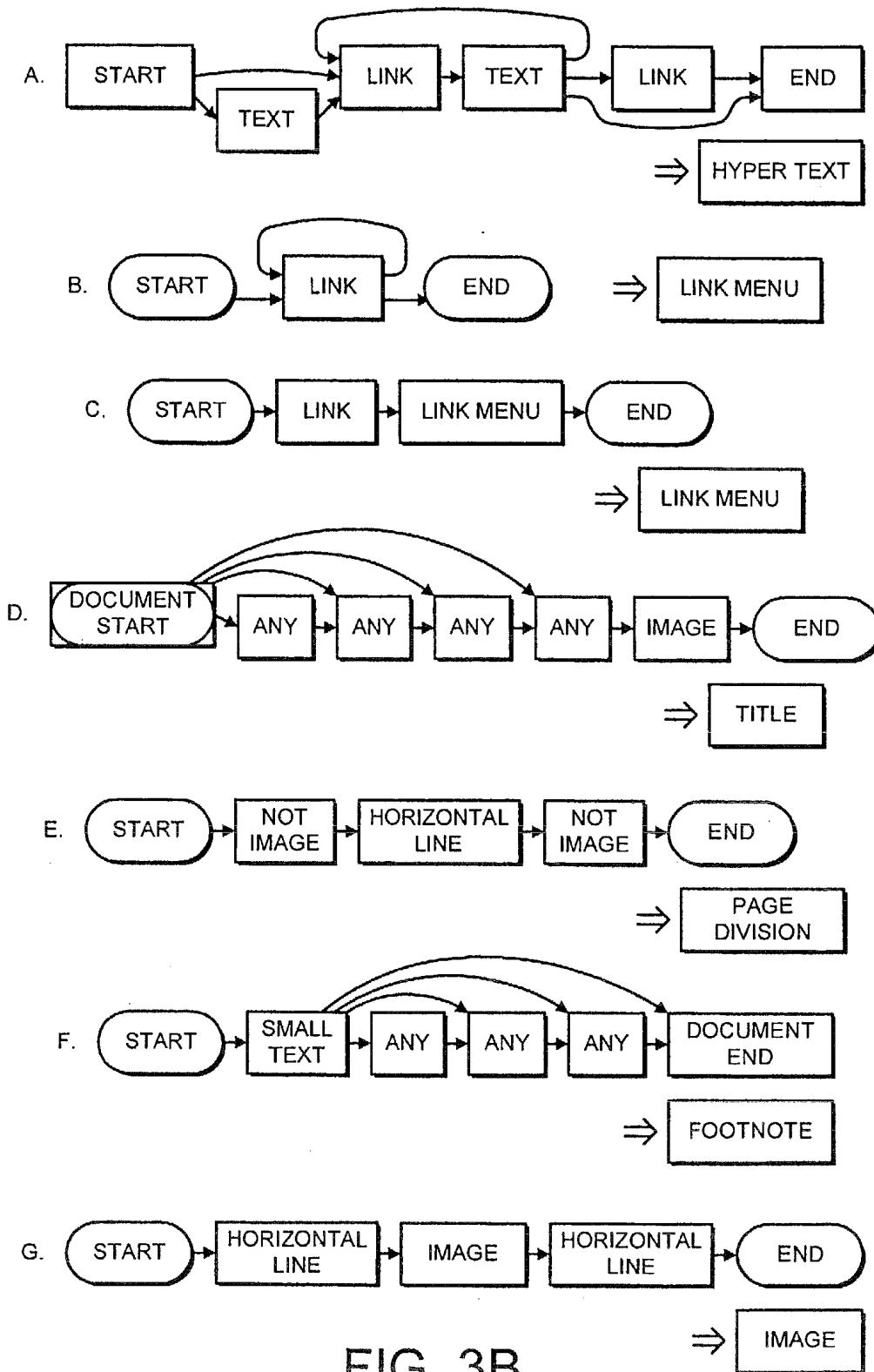


FIG. 3B

05 + 3 01

4 / 6

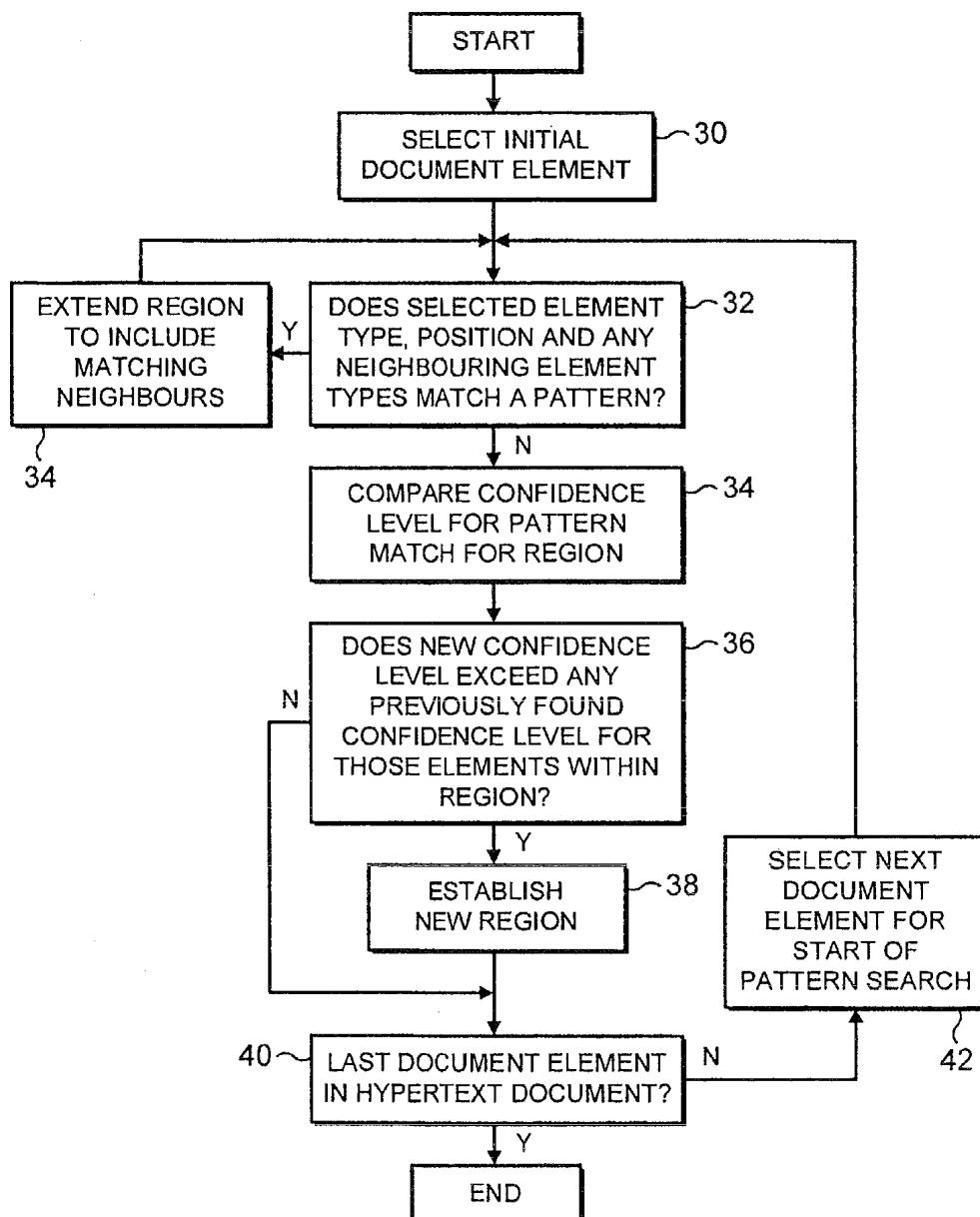


FIG. 4

05 + 3 01

5 / 6

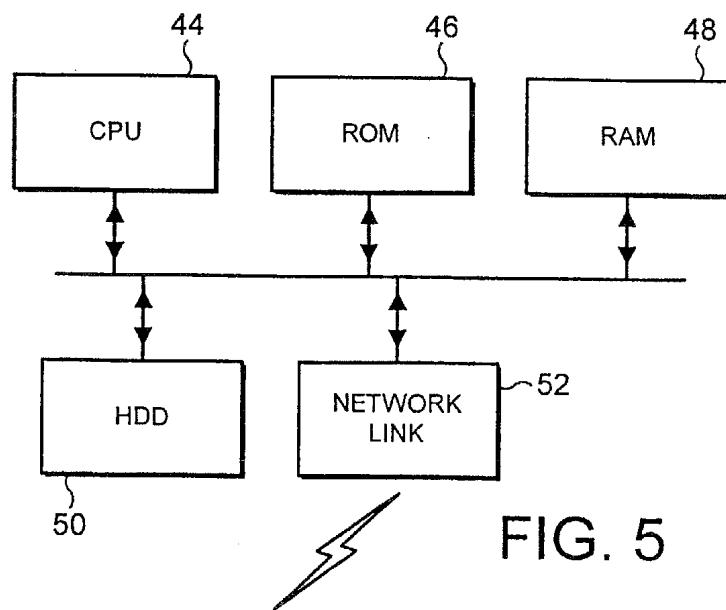


FIG. 5

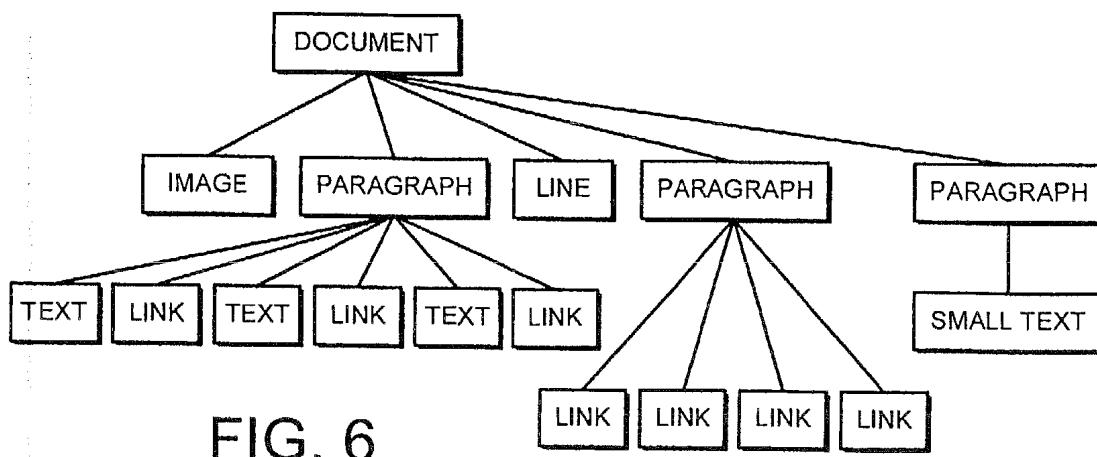


FIG. 6

05 3 01

6 / 6

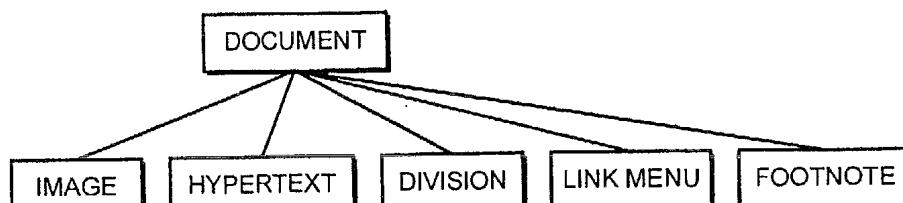


FIG. 7

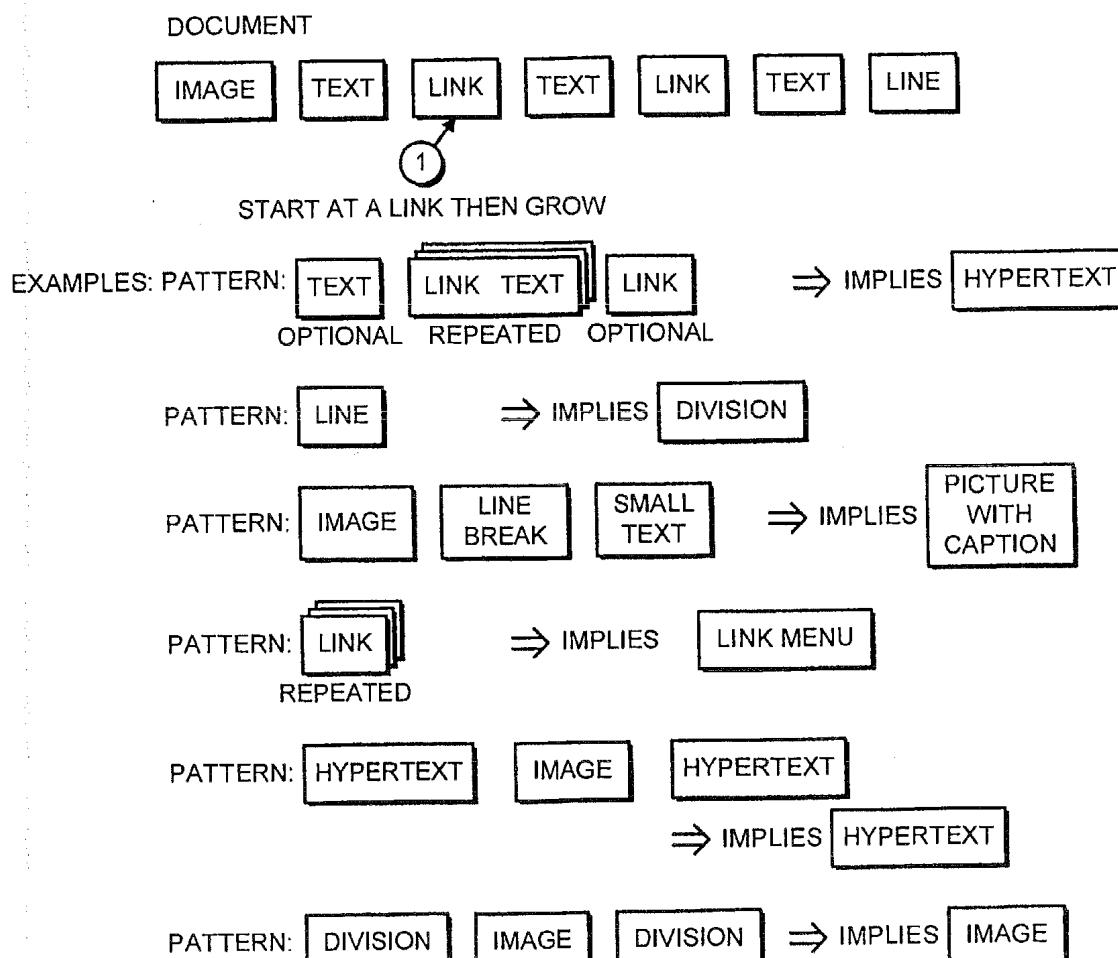


FIG. 8

**2366633**

1

ANALYSING HYPERTEXT DOCUMENTS

This invention relates to the field of the analysis of hypertext documents. More particularly, but not exclusively, this invention relates to the type of analysis of hypertext documents that is useful in transcoding such documents from a form suitable for display on a conventional personal computer to a form suitable for display on a mobile telephone or other display device having a more limited display capability.

Internet web page transcoders attempt to reduce page content, summarise, divide, or otherwise process a web page on the basis of its content with the aim of making the material more suitable for display upon devices for which that material was not originally intended. Currently, transcoding systems operate using very localised HTML element processing. However, there is a limit to the degree of sophistication in the processing that can be performed at an individual element level and this impedes the improvement in intelligence and functionality of transcoding systems.

Some transcoders operate as a "proxy" on a computer on a network which can be contacted by the end user's device in place of it contacting the computer holding the desired document. Other transcoding systems can operate as "plug-ins" to computer software which dispatches documents from the original computer. It is also possible that the transcoding system could reside on the user's computer that receives the document.

Viewed from one aspect the present invention provides a method of identifying one or more regions of a hypertext document formed of a plurality of document elements, said method comprising the steps of:

parsing said plurality of document elements to identify an element type for each of said plurality of document elements;

25 categorising one or more patterns of element types within said hypertext document indicative of respective categorised regions of said hypertext document; and

selecting one or more categorised regions to be identified as said one or more regions of said hypertext document in dependence upon a confidence measure associated with each categorised region.

30 The invention recognises that a more abstract view of a web page being processed can be used to identify the structure of that web page in a manner that assists

subsequent processing. In particular, the invention steps away from the analysis of individual elements and instead parses the individual elements for element type and looks for patterns amongst those elements types. This allows different regions of a hypertext document to be identified in a way that was not previously possible.

5 Whilst the identification of different regions of a hypertext document may be useful for various different reasons, it is particularly advantageous when it is desired to divide an import hypertext document into smaller processed documents as the invention allows more intelligent and usable divisions between the processed documents to be achieved.

10 Whilst strictly deterministic algorithms could be used for the pattern matching, it is preferable to use heuristic algorithms as these are more able to cope with the wide variety of different hypertext document layouts that can be encountered whilst still accurately identifying the different regions of a document.

15 Whilst the pattern matching used could take a variety of forms, a particularly efficient form of pattern matching is one in which each document element is compared with its neighbours to see if there is a consistent pattern of relationship between them at which point these matching elements may be grouped together and additional new neighbours considered. Accordingly, the group of elements matching a given pattern will grow by sequentially including neighbouring elements until neighbouring 20 elements are encountered that no longer fit the pattern.

Additionally, patterns between neighbouring or non-adjacent regions may also be subsequently matched.

25 A further characteristic of document element that has been found useful in identifying document regions is the relative position of a document element within a hypertext document as a whole. As an example, images at the top of a document are often associated with a title of a document and small text at the bottom of a document is often associated with a footnote.

30 The technique of the present invention has been found to be particularly effective in identifying mixed blocks of text and links as hypertext prose, blocks of adjacent neighbouring links as link menus, text neighbouring a link menu as a title for the link menu, an image close to the top of a document as part of a title, small text

toward the bottom of a document as part of a footnote and a horizontal line spaced away from any images or graphics elements as a division within a document.

It will be appreciated that the technique of the present invention may be used in the analysis of many different types of hypertext document, e.g. documents containing 5 both text and additional elements such as images, graphics and links. The invention is particularly well suited to the analysis of HTML, WML and CHTML documents but could be applied to other hypertext document formats such WML documents or even PDF documents and the like.

Viewed from another aspect, the present invention provides apparatus for 10 identifying portions of a hypertext document formed of a plurality of document elements, said apparatus comprising:

parsing logic operable to categorise said plurality of document elements to identify an element type for each of said plurality of document elements;

categorising logic operable to identify one or more patterns of element types 15 within said hypertext document indicative of respective regions of said hypertext document; and

selecting logic operable to select one or more categorised region to be identified as said one or more regions of said hypertext document in dependence upon a confidence measure associated with each categorised region.

20 The invention also provides a computer programme storage medium for storing a computer programme to control a general purpose computer to operate in accordance with the above described techniques. The computer storage medium could be a CD, a hard disk drive or a downloaded computer file.

An embodiment of the invention will now be described, by way of example 25 only, with reference to the accompanying drawings in which:

Figure 1 schematically illustrates the technique of the present invention in analysing a hypertext document;

Figure 2 is a simplified flow diagram illustrating the processing performed in the technique in Figure 1;

30 Figures 3A and 3B are a table indicating the correspondence between patterns of element types, element positions, and the portions of a document to which these correspond and a "regular expression" view of the same relationships;

Figure 4 is a flow diagram illustrating in more detail how the pattern matching may be performed;

Figure 5 is a schematic illustration of a data processing apparatus that may be used to implement the techniques of the present invention;

5 Figures 6 and 7 illustrate a source hypertext document viewed at element type and region levels; and

Figure 8 illustrates various example patterns which may be matched.

10 Figure 1 schematically illustrates the technique of one example of the present invention. An input hypertext document in the form of a HTML page 2 is the starting point for the process. This HTML page 2 has been designed by its author to be suited for display upon a computer monitor 4 of a typical personal computer. If it is desired to access this HTML page, or at least the content of that page, using a mobile telephone 6, or other small display screen device such as a personal digital assistant, then the original form of the HTML document 2 is inappropriate for display upon the 15 mobile telephone 6. More particularly, the mobile telephone 6 will have a small and less capable display than is provided by the computer monitor 4 and accordingly only a small portion of the contents of the HTML page 2 could be fully displayed at any given time.

20 For this reason, transcoding products seek to reduce, sub-divide and simplify the content of source pages in a manner that enables them to be better displayed upon less capable (or just different) display devices. The transcoding processing may take place in a proxy server disposed between the user device and the source server or it may take place in the source server itself.

25 The first step in the analysis of the HTML document is to identify the different element types within that document. Accordingly, the image 8 is identified as an image type 8', the text 10 is identified as a text type 10', the link 12 is identified as a link type 12' and so forth. The horizontal line 14 is identified as a horizontal line type 14' and the footnote text 16 is identified as a small text type 16'. Thus, the first step in the processing of the HTML document achieves an indication of each document 30 element as a particular element type. This might be represented as a "parse tree" as in Figure 6.

The next stage in the processing is to perform pattern matching upon the element types to find different regions of the original HTML document 2 (this may use heuristic algorithms). In particular, the image 8 and its associated image type 8' positioned towards the top of the HTML document 2 is identified as part of a title portion. Any text immediately preceding the image 8 or following the image 8 and spaced from other content within the HTML page 2 may also be grouped together with the image 8 and identified as part of the title.

The mixed section of text 10 and links 12 is pattern matched to a block of hypertext prose. Such pattern matching may be achieved by starting with the initial 10 document element and then comparing it with its neighbours to identify a pattern consistent with a particular type of document portion and growing that portion by encompassing further neighbouring elements until the identified pattern no longer holds true. A block of hypertext prose will typically form the main prose of an HTML page and represent highly significant information content. Accordingly, in a 15 transcoding system such a portion may be identified as the most significant to a user and presented first to that user. Content containing elements rather than layout elements are given a greater weighting in assessing user importance.

The horizontal line 14 within the HTML page 2 is identified as a horizontal line type 14'. As this horizontal line type 14' is spaced apart from any images or 20 graphical images within the HTML page 2, it is pattern matched to represent a division within the HTML page 2. It has been found that if a horizontal line is close to or neighbouring an image element or a graphical element, then it is more usually intended by the author of the HTML page 2 to form part of that image or graphic (e.g. part of the border) and accordingly if such a disposition is detected then the horizontal line 25 will not be pattern matched to represent a division.

Tags indicative of the document structure per se (e.g. paragraph tags) are given their original meaning in deciding how to divide up a page.

A sequence of adjacent hypertext links 18 is pattern matched to represent a link menu. If such a link menu is closely preceded or followed by a text element, then that 30 text element is taken to represent a title or footnote to the link menu.

Small text identified towards the bottom of a hypertext page 2 or close to an image 8 is identified as being a footnote region or caption . The confidence in the

identification of such small text as a footnote region is determined as higher if the small text is spaced from preceding elements within the HTML page 2.

Once all the different document regions have been identified, then the original content of the HTML page 2 may be divided into processed documents (cards) that 5 may be separately supplied to and displayed by a mobile telephone 6. A division (or structural markup element) identified within the HTML page 2 is taken as a strong indication for a point at which the original content can be divided. The order in which the processed documents will be displayed to a user may be selected in dependence upon the nature of the document portion that has been pattern matched, e.g. a main text 10 of mixed hypertext prose and a link menu will be regarded as more significant and more desirable to display to a user than would be a title page or a footnote page.

The patterns matched may be indicative of more than one way in which the document may be divided into regions. All possible ways in which patterns found may 15 identify regions are recorded with associated confidence measures. When the patterns have all been matched then the regions to be used are selected as those having the highest confidence measures . A given document element may lie in two or more possible regions, but these highest confidence region will be used.

Figure 2 is a simplified flow diagram of the processing that may be performed as part of a transcoding process utilising the present invention. At step 20 a web page 20 is received by the system. At step 22 the web page is parsed to identify different document elements within it. The markup language form of a web page makes it relatively straight forward to identify different document elements by their associated tags. At step 24 each of the identified document elements is noted as a particular type 25 of element, for example the HTML <font size = 2> note </font> would be identified as "small text". At step 26 a search is made through the element types to identify patterns characteristic of different regions of a document. This pattern matching may use heuristic algorithms with associated confidence levels for the pattern matching achieved in accordance with known pattern matching techniques. At step 28, the received web page is divided into a number of separate processed documents in dependence upon the identified document regions with the highest confidence 30 measures as produced by step 26.

5 The division on an input web page in accordance with the present technique may be coupled with other transcoding techniques to reduce the content of a web page to focus upon the content useful to a user and adapt the web page to a form more suited for manipulation using a device other than that for which the web page was originally designed. The technique of the present invention is strongly advantageous as part of a transcoding system as a whole.

10 Figure 3A is a table indicating a relationship between element types, element positions and corresponding regions. The pattern matching discussed above is responsive to the element types and element positions to heuristically identify corresponding document regions. The pattern matching performed may associate a confidence level with each pattern matched. Accordingly, a particular document element might form part of two possible matching patterns indicative of different document portions. The confidence levels associated with the pattern matches can be used to select which of the pattern matches forms the basis of subsequent processing 15 and the division of the document in accordance with the identified regions. Patterns between identified regions may also be pattern matched to identify larger regions.

20 Figure 3B corresponds to Figure 3A, but in this case gives a regular expression view of the relationships between elements type that may be pattern matched. Where several exits are shown from a stage, any may be taken depending upon the next element type to be processed with the source document.

25 Figure 4 is a flow diagram schematically illustrating the processing performed in the pattern matching. At step 30, an initial document element within an input HTML page is selected. This will typically be the first document element in the page. Step 32 compares the selected element type, position and any neighbouring element types to determine if a match to one of a plurality of predetermined patterns is indicated with a sufficient degree of confidence. If such a match is indicated, then processing proceeds to step 34 at which an attempt is made to extend the region to include further neighbouring elements that would also match. Accordingly, the region of a document matching a given pattern is grown by successive processing in step 32 30 and 34 until the matching region can be extended no further. At this stage processing proceeds to step 34.

Step 34 checks the confidence level associated with the pattern match achieved in the preceding steps to determine whether this has achieved identification of a document region with a greater or lesser degree of confidence than any preceding match identified for the elements within that document region. Steps 36 and 38 serve 5 to either establish the newly matched region as the current preferred candidate or not in dependence upon the comparison of the confidence levels at step 34.

Step 40 checks to see whether the last document element within the hypertext document has already been used as a starting point for the pattern matching process. If this is not the case, then processing proceeds via step 42 at which the next document 10 element is selected as a starting point than that which previously formed the starting point for the entry into the pattern matching steps 32 and 34. Thus, each document element in term is used as a starting point for pattern matching until all of the document elements have been so used. The highest confidence level patterns and portions resulting when the processing has finished are those used for subsequent 15 transcoding operations, such as page division. Subsequently, identified regions may be used as the starting point for further region/region or region/element matching.

Figure 5 schematically illustrates a data processing system of the type which may perform the technique of the present invention. This data processing system includes a central processing unit 44, a read only memory 46, a random access 20 memory 48, a hard disk drive 50 and a network link 52. Such a general purpose data processing system will execute a computer program that may be stored upon the hard disk drive 50, within the read only memory 46 or downloaded via the network link 52. The working memory during such program execution will be provided by the random access memory 48. The results of such data processing may be displayed to a user of 25 another device with which the system communicates through the network link 52. The user can give commands to the system via the user input/output unit 58 in conjunction with the keyboard 60 and the mouse 62. It will be appreciated that the central processing unit 44 executing computer program instructions effectively serves as logic for performing the processing steps described above. The computer program executed 30 by the data processing system may be loaded into the system via a tangible medium, such as a compact disk or floppy disk, or downloaded via the network link 52.

Figure 6 illustrates the structure of the document of Figure 1 in the form of the element types identified and the structural markup tags that may be included. Figure 7 is a view of the same page at a higher level of abstraction once the document regions have been identified.

5. Figure 8 illustrates various document element and region patterns that may be matched to predefined criteria indicative of certain document regions. Many further patterns are possible.

CLAIMS

1. A method of identifying one or more regions of a hypertext document formed of a plurality of document elements, said method comprising the steps of:
  - 5        parsing said plurality of document elements to identify an element type for each of said plurality of document elements;
  - 10        categorising one or more patterns of element types within said hypertext document indicative of respective categorised regions of said hypertext document; and
  - 15        selecting one or more categorised regions to be identified as said one or more regions of said hypertext document in dependence upon a confidence measure associated with each categorised region.
2. A method as claimed in claim 1, further comprising the step of dividing said hypertext document into a plurality of processed documents in dependence upon said regions of said hypertext document identified by said one or more patterns of element types.
3. A method as claimed in claim 2, wherein different regions of said hypertext document appear in respective different processed documents.
- 20        4. A method as claimed in any one of the preceding claims, wherein said step of categorising one or more patterns uses heuristic algorithms.
- 25        5. A method as claimed in any one of the preceding claims, wherein said step of categorising one or more patterns starts with each document element and compares an element type for that document element with element types for neighbouring document elements to identify a pattern of neighbouring element types indicative of a region of said hypertext document.
- 30        6. A method as claimed in claim 5, wherein said step of categorising one or more patterns starts with each of any hypertext links within said hypertext document.

7. A method as claimed in claim 5, wherein said step of categorising one or more patterns starts with each of any textual headings within said hypertext document.
8. A method as claimed in claim 5, wherein said step of categorising one or more patterns starts with each of any graphical images within said hypertext document.
9. A method as claimed in any one of claims 5 to 8, wherein said comparison is made with previously categorised regions of said hypertext document.
10. 10. A method as claimed in claim 9, wherein said comparison of categorised regions is made between non-consecutive regions of said hypertext document.
11. 11. A method as claimed in any one of the preceding claims, wherein said step of categorising one or more patterns combines one or more of the methods claims 5, 6, 7 and 8.
12. 12. A method as claimed in any one of the preceding claims, wherein said step of categorising one or more patterns is responsive to a relative position of a document element within said hypertext document as a whole when identifying a portion of said hypertext document.
13. 13. A method as claimed in any one of the preceding claims, wherein a plurality of possible categorised regions may correspond to a given document element and a confidence measure is associated with each of said plurality of possible categorised regions.
14. 14. A method as claimed in claim 13, wherein, after all overlapping possible categorised regions have been found, overlapping categorised regions are compared and a categorised region having a highest confidence level is selected from amongst other categorised regions that overlap its document elements.

15. A method as claimed in any one of the preceding claims, wherein a mixed block of text and links is identified as hypertext region.

5 16. A method as claimed in any one of the preceding claims, wherein a plurality of neighbouring links is identified as a link menu region.

17. A method as claimed in claim 16, wherein text immediately preceding said link menu region is identified as a title for said link menu region.

10 18. A method as claimed in any one of the preceding claims, wherein an image close to the top of said hypertext document is identified as part of a title portion.

15 19. A method as claimed in any one of the preceding claims, wherein text close to the bottom of said hypertext document and with a font size smaller than normal for said hypertext document is identified as part of a footnote portion.

20 20. A method as claimed in any one of the preceding claims, wherein a horizontal line spaced away from any images or graphics within said hypertext document is identified as a division.

21. A method as claimed in any one of the preceding claims, wherein a textual heading is identified as a division.

25 22. A method as claimed in any one of the preceding claims, wherein a boundary between table cells is identified as a division.

23. A method as claimed in claim 2 and any one of claims 20 to 22, wherein said division is used to identify where said hypertext document should be split into processed documents.

24. A method as claimed in any one of the preceding claims, wherein a structural markup element is recognised as its original meaning.

5 25. A method as claimed in any one of the preceding claims wherein a value of a relative importance to a user of each categorised region is allocated for later processing.

10 26. A method as claimed in any one of the preceding claims, wherein said method of categorising regions is part of a method for converting a hypertext document originally intended for display on a first type of display device into a form adapted for display on a second type of display device.

15 27. A method as claimed in claim 26, wherein said first type of display device is a conventional computer monitor.

20 28. A method as claimed in any one of claims 26 and 27, wherein said second type of display device is one of: a mobile telephone display and a personal digital assistant display.

25 29. A method as claimed in any one of the preceding claims, wherein said hypertext document is an HTML document.

30. A method as claimed in any one of claims 1 to 29, wherein said hypertext document is one of a WML document and a CHTML document.

30 31. A method as claimed in any one of the preceding claims, further comprising the steps of employing iterative growth and combination of categorised regions of said document so as to reach one of a preferred number of said regions and a preferred size of each said region.

32. A method as claimed in any one of the preceding claims, further comprising the step of dividing the said hypertext document into a plurality of processed documents in dependence upon a size of said categorised regions.

5 33. A method as claimed in any one of the preceding claims, further comprising the step of dividing the said hypertext document into a plurality of processed documents in dependence upon a measure of importance to a user of said categorised regions.

10 34. A method as claimed in any one of the preceding claims, further comprising the step of dividing the said hypertext document into a plurality of processed documents in dependence upon a predetermined maximum size for said plurality of processed documents matched to a target display device size.

15 35. A method as claimed in any one of the preceding claims, wherein said method steps are performed in real time upon an access request to said hypertext document.

36. A method as claimed in any one of the preceding claims, wherein said steps of parsing, categorising and selecting are performed on a computer acting as a proxy 20 between a user device requesting said hypertext document and a server storing said hypertext document.

37. A method as claimed in any one of claims 1 to 34, wherein said steps of parsing, categorising and selecting are performed on a server storing said hypertext 25 document prior to dispatch of a modified form of said hypertext document to a user device.

38. Apparatus for identifying portions of a hypertext document formed of a plurality of document elements, said apparatus comprising:

30 parsing logic operable to categorise said plurality of document elements to identify an element type for each of said plurality of document elements;

categorising logic operable to identify one or more patterns of element types within said hypertext document indicative of respective regions of said hypertext document; and

5. selecting logic operable to select one or more categorised region to be identified as said one or more regions of said hypertext document in dependence upon a confidence measure associated with each categorised region.

39. A computer program storage medium for storing a computer program to control a computer to perform a method as claimed in any one of claims 1 to 36.

10

40. A method of identifying portions of a hypertext document substantially as hereinbefore described with reference to the accompanying drawings.

15

41. Apparatus for identifying portions of a hypertext document substantially as hereinbefore described with reference to the accompanying drawings.

20

42. A computer program storage medium for storing a computer program to control a computer to perform a method substantially as hereinbefore described with reference to the accompanying drawings.



INVESTOR IN PEOPLE

Application No: GB 0021881.8  
 Claims searched: 1 - 42

Examiner: Natasha Chick  
 Date of search: 8 June 2001

**Patents Act 1977**  
**Search Report under Section 17**

**Databases searched:**

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.S): G4A AUDB AUXX

Int Cl (Ed.7): G06F 17/21 17/30

Other: Online: EPODOC, WPI, PAJ, IEL

**Documents considered to be relevant:**

Category	Identity of document and relevant passage	Relevant to claims
A	EP 0994426 A2 SAMSUNG ELECTRONICS	
A	EP 0969389 A2 IBM	
A	Wireless-adaptation of WWW content over CDMA, Ham et al. Mobile Multimedia Communications 1999. Pages 368-372	

X Document indicating lack of novelty or inventive step	A Document indicating technological background and/or state of the art.
Y Document indicating lack of inventive step if combined with one or more other documents of same category.	P Document published on or after the declared priority date but before the filing date of this invention.
& Member of the same patent family	E Patent document published on or after, but with priority date earlier than, the filing date of this application.